

2017 July 9 周报

Junhua Lu

Done

1. Minor revisions on BeXplorer (sec 4, model).
2. Finish the patent for netease company. Prepare for the next (律师称正在修改中, 随后会联系我)
3. Almost done in data cleaning. Discussion on data extraction and model.
4. Push forward the application of DS2019 for US J-1 visa.
5. Begin revision of BeXplorer. Minor drafts made. Some JS techniques learned.

To Do

2. Model testing and paper revision for BeXplorer.

论文阅读:

WWW 2015 LINE: Large-scale Information Network Embedding 上次说的网络嵌入的一种, 考虑的是一阶(直接连接, 有强联系距离越近)和二阶(有很多共同连接距离越近)两种 proximity. 细节的说, 对于点 u, v 和他们的边 (u, v) , 一阶 proximity 权 w_{uv} ; 二阶, 首先令 $p_u = (w_{u,1}, \dots, w_{u,|V|})$ 为 u 和所有其他节点的一阶 proximity, 那么 u 和 v 二阶 proximity 就是 p_u 和 p_v 之间的相似性. 这里二阶用了 softmax, 常用于多标签预测, 和概率分布相结合, 其用于神经网络计算时由于函数特性使得计算非常方便.

ACL 2017 (annual meeting of the Association for Computational Linguistics) CANE: Context-Aware Network Embedding for Relation Modeling 文章做的问题是从网络结构中来探索关系, 比如根据你 facebook 上的社交网来判断两个人之间可能是什么关系, 做法是从网络 embedding 的角度出发. 但是传统的 embedding 都会面临问题就是每个节点的 embedding 都是固定的, 然而实际情况中节点与不同邻居交互时会产生不同的面貌. 举例来说, 一个 NLP 教授与一个教授合作一篇文章, 他们这篇文章合作的子方向, 与该教授与另一个人合作的文章的子方向并不一致, 就是说在不同 context 下有不同的角色. 本文提出的方法, 从两个角度, 一个是传统的结构角度搞 embedding 的优化(类似前面说的 LINE), 一个是用基于文本的信息网络得到上下文搞一个 embedding. 这两个 embedding 为了让他们可以兼容, 在做文本角度 embedding 时候, 映射到了同一个表示空间. 这里文本角度, 又分为了上下文无关于上下文有关的方法, 前者只要来个 CNN 就行; 后者定义了一个叫 mutual attention 的矩阵, 在他们的词序列中间加了个 attentive 矩阵, 然后分别做 pooling. 最后用负采样来优化速度. 虽然感觉 KDD 上会有很多文章都是像谜一样的从一个地方预测另一个, 但本文从我角度看, 可能具有比较大的普适性, 因为被泛化为网络结构与上下文信息的综合体.

EuroVis 2017 NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations Named-Entity Recognition 中文叫命名实体识别, 是 NLP 中一个比较基础的任务, 其目标在于识别出文本之中人名、地名等专有名词和有意义的时间、日期等数量短语(所谓实体)并加以归类 or 分类. 识别出这些以后, 对于文本的分析会很有帮助. 本文从个人感觉来说应该是深刻结合了用户需求做的一个可视化系统, 从那些党派之间辩论的数据(verbatim conversational transcript, 逐字逐句记录的, 因为也是应急发挥的所以可能不同于书面文章)出发探索各种领域专家想要做的任务. 其设计很简单, 就是在力引导的节点链接图上做花样. 其视图间是联动的, 从层级上说是依据分析任务不断深入(抽象);

其中有个 anchored 力引导, 在展示时候, 对某些类型实体点固定, 别的点去自由的动, 便于分析彼此之间关系. 最后加上了时间轴及对应编码, 使得信息更完备.

VIS 2015 *A Visual Voting Framework for Weather Forecast Calibration* 也是类似 ensemble 可视化的方法, 去年的 nested PCP 有些编码和这篇十分接近. 这里的 visual voting 借鉴了类似集成学习的思想, 只是针对天气预报数据及方法将其可视化化了, 把人的交互恰到好处的融入到其中.

WSDM '17 Proceedings of the Tenth ACM International Conference on Web Search and Data Mining *Motifs in Temporal Networks* 本文主要工作是将动态网路中一些小的结构特性(星形\clique\三角等等, 一般就是 induced subgraph, 成为 motif)及其时序特征的寻找和加速算法. 这里动态网路的体现在于边是带时间戳的, 本文动态 motif 还限定了时间长度. 文章限定了时序特征在于计数某些特征的数量上.